Unstructured Data: What It Is and How It Is Used

Lauren Dassinger

Department of Information Systems, Middle Tennessee State University

INFS 6790: Seminar in Database

Dr. Stoney Brooks

October 2, 2022

Abstract

This paper examines unstructured data and how it is used in business. Understanding how unstructured data plays a significant role in the industry will allow missed opportunities to become future innovations. Unstructured data is data that cannot be contained within rows and columns. The amount of utilization that unstructured data can carry in a business is abundant. When it comes to finding trends and paving the way for future endeavors, unstructured data will become the foundation. This paper will explore where it originates from, its pros and cons, how it compares to structured data, and how it is being used in analytics and machine learning.

Keyword: Unstructured data, unstructured data mining, unstructured data machine learning, unstructured data analytics

Unstructured Data: What It Is and How It Is Used

Data is transformed into information that all organizations rely on daily. Obtaining this data on current and future customers allow organizations enough information to meet customer needs. Unstructured data is used to collect as much information from one customer in all different ways. It has a versatile structure that can be used toward focusing on a demographic market thus giving an organization the ability to fulfill open needs. This paper outlines how unstructured data is and how it is used in analytics. It will break down the importance of the advantages and the disadvantages that comes along with it. The paper's focus will state the proper ways of understanding so that when an organization adapts to how unstructured data can reach its full utilization. By understanding these key concepts, the organization can excel using unstructured data in analytics, and innovative future uses.

Unstructured data is defined as data that does not conform to a data model and has no simply recognizable structures (Ihritik, 2021). Unstructured data include data that cannot conform to database rows and columns. It does not follow any semantics or rules, lacks format or sequence, and cannot be used in computer programs easily due to not containing an identifiable structure (Ihritik, 2021). The unstructured data source comes from web pages, images, social media, and more.

Examples of Unstructured Data

Web pages are unstructured data due to the containment of text, images, videos, forms, and functionalities. HTML is a part of web pages, yet this code cannot depict the meaning of the content within the page when a business needs to analyze the content data to view customer behavior, competitors, and opportunities (Rose, 2021). Acquiring this level of information makes it challenging to know the potential web pages without breaking them down into unstructured data. Enabling this ability will allow an analyst to go further than the content on the page and the code listed.

Images are unstructured data because databases do not process or identify the actual contents of the media file (Rose, 2021). Obtaining images becomes a resource for being analyzed. It has the potential to find trends that may not have been seen. An example of images that are unstructured are TIFF, JPEG, GIF, PNG, or RAW. The contents of these images are consisted of their own characteristics.

Social Media are unstructured data by use of the text within a post. In this context, when an individual writes about their day and then posts it, there is no way to categorize it into a row or column. Other aspects of social media do not have structured data, for example, information on followers, friendships, groups, or networks (Rose, 2021). Yet the value of the post has the potential to investigate their interest. An example could be where they frequently eat or drink coffee to what items they will likely purchase.

Advantages of Unstructured Data

Many components make unstructured data beneficial within an organization. The two main benefits of using unstructured data within an organization are to improve customer experience and to advance innovation. Unstructured data can enhance the customer experience when an organization captures it to identify the customer demographic. Collecting the unstructured data of social media platforms and then finding the trends as to the individual who purchases the product being sold. This way, the organization can have a foothold in the competition, which is how they advance innovation. Staying on top of the trends enables them to see where the gaps are and close in on them (Davies, 2021).

Another advantage is the data compact. Meaning it does not solely algin in text form, which makes it easier to manipulate and use the pieces you need instead of the entirety that happens in structured data. Unstructured data is more commonly used in several business intelligence applications or tools, such as Azure, Hadoop, Accenture, and more (Hansen, 2022). The following list the main advantages:

- Supports the data which lacks a proper format or sequence
- A fixed schema does not constrain data

- Very Flexible due to the absence of a schema
- Data is portable
- Very scalable
- Easily with the heterogeneity of sources
- Variety of business intelligence and analytics applications (Ihritik, 2021)

Disadvantages of Unstructured Data

On the other hand, unstructured data can have its downside. One of its significant shortcomings is storage since unstructured data cannot properly store in the standard way of storage since data cannot contain within rows and columns. Other disadvantages include analyzation, specialized tools (these tools require extra training rather than normalized tools like Tableau), and indexing. With the specialized tool come with training. Majority of analyst are not trained in having to use the tools that are needed to transfer the raw data into information. Having to analyze unstructured data is a challenge if it is being implemented into an organization that does not have people trained and is not equipped with the proper tools to analyze it. As for indexing, there is a higher probability that errors will be frequent, and attributes will lack pre-defined (2021, July 8).

Unstructured Data versus Structured Data

How are structured data and unstructured data so different? Structured data align within rows and columns, have numbers and dates, and requires less storage. Structured Data have fields that store data like people's first names, last names, or addresses. Text strings are inside records which makes it simple to search for what to look up. This data can be human or machine-generated for as long as it is created in a rational database. In a rational database, the format is searchable by algorithms or human-generated queries applying types of field names and data. Examples comprise of date, numeric, or currency, as well as alphabetical (Taylor, 2022). The most common way to manage structured data stored in databases uses SQL (Ihritik, 2021). Some structured characteristics that are only inside this type of structures include:

- Data conform to a data model and have an easily identifiable structure
- Data is stored in the form of rows and columns
- Data is well organized, so, Definition, Format, and Meaning of data are known
- Data resides in fixed fields within a record or file
- Similar entities are grouped to form relations or classes
- Entities in the same group have the exact attributes
- Easy to access and query so that other programs can easily use that data
- Data elements are addressable, which makes it efficient to analyze and process (Ihritik, 2021)

Unstructured data, as seen previously, includes audio and word processing files and requires more storage. Storage needs to be more prominent in unstructured data than structured data due to the high volume of metadata unstructured data produces. This data comprises mainly of written various arrays. However, this does not mean that the data lacks all structure. Data is conformed in some ways. An instance is interview data; it is known that the questions asked to have influenced the answers given (Boulton, & Hammersley, 2006). Typically, in interview data it comprises of questions, answers, notes, and other methods. This is not structured in that there are no common responses. Each person will answer differently based on the same question. It benefits analysis in that each person may answer differently, yet their core foundation is the same, and making this unstructured data. Breaking down this thought process helps one to understand how the interview is broken down into pieces that are needed to be analyzed. (see Appendix A for more information on how structured and unstructured data compare). The foundation has been set on unstructured data and now is how this data is interpreted using analytics.

Analytics of Unstructured Data

Analytics is defined as viewing data as it coordinates the business needs. This way of thinking combines raw data and transforms it into information that helps the business achieve its objectives. Analytics can be historical or predictive trends that help a business focus on where the budget and other planning need attention. There are many types of job roles that fall into this category of analytics. The majority of them are either business analysist or business intelligence. Both of these job roles use unstructured data to perform their analysis. Tools that are used to transform unstructured data into models are platforms such as Power BI and Tableau. The tools mentioned are used to find trends by applying big data analysis, qualitative analysis, data mining, and data modeling. Modeling unstructured data can be difficult when attempting to use for the first time. Here are some tips when analyzing unstructured data:

- Select the End Goal
- Select a Method to Analyze
- Have the Sources of all Data Identified
- Evaluate the Technology
- Have Real-Time Access
- Use Data Lakes
- Clean the Data
- Recover, Classify and Segment Data
- Use Visualization (2019, May 28)

Big Data

What is considered big data, and how can it be analyzed within unstructured data? Big data is a large amount of data that contains Volume, Velocity, Variety, and Value (4Vs) (What is Big Data? Big Data explained). Having this amount of data can be overwhelming to look at and analyze. Tools are available to help people break apart when searching for or needing to solve a problem. Unstructured data comes into play when businesses must know data to have data support activities.

An example is a large company collecting data on its clients to observe their shopping habits. This data could have different formats, such as images. When this type of data is stored, it is stored as an object or No-SQL database, thus collecting over time. Even though it is challenging to analyze and process unstructured data, it provides a greater depth of analysis (Chasupa, & Paireekreng, 2021). (See Appendix B on unstructured big data extracting model and trial run).

Qualitative Analysis

The objective of the qualitative analysis is to use judgment to understand the company's value or prospects based on non-quantifiable information. This can be industry cycles, management knowledge, research strengths and development, and labor associations (Smith, 2022). When unsure of what to expect, one needs to define the problem or develop an approach to the problem (Botha, 2019). Unstructured data is used within this process to subjective opinions and judgment of the brand in the form or text. An example of how using qualitative analysis with unstructured data allows the business to get a sense of different demographics and age groups.

Data Mining

Why is mining unstructured data valuable? Mining data comprises adding up the values in a data set, then dividing by the number of values that have been added. It generally involves using techniques to analyze and improve business processes using data from multiple sources. This process transforms raw data into usable information through charts and other visualization tools (2022, February 25). Mining unstructured data (MUD) techniques have a variety of helpful intel. One approach has been used to summarize bug reports. The focus is to identify and extract the most relevant sentences of bug reports (a specific report that outlines information about what is wrong and needs fixing with software or on a website) by using supervised learning techniques, network analysis, and information retrieval (Bavota, 2016). When accessing MUD techniques, a few steps must be kept in mind before and after extracting data. They include seven steps that most analysts go by:

1. Data cleaning to align data with the industry standards
2. Data integration to combine data sets and sources
3. Extract relevant information to analyze and evaluate
4. Transform and consolidate data to prepare for mining
5. Data mining
6. Evaluate data and pinpoint patterns
7. Create data reports to share information with appropriate parties (Bavota, 2016).

Data Model

Data modeling is creating a simplified diagram of a software system and the data elements it contains, using text and symbols to represent the data and how it flows. Data models provide an outline for how the database needs to be designed. How unstructured data fits into data modeling is that conventionally it does not. Unstructured data is best managed in No-SQL. The other possibility to collect unstructured data is to use a data lake to preserve it in raw form (2021, June 29). No-SQL works with unstructured data because it can store it without the need for the data to have rows and columns. A data lake is a centralized repository designed to store, process, and secure large amounts of different kinds of data, that includes unstructured data. Data lakes are stored without organization or hierarchy that allows raw data in its current format.

Obtaining the understanding of how unstructured data and analytics correlates with qualitative and big data analysis, data mining, and data modeling, the next question is, what's next? How can analyzing unstructured data become a highly valued resource that most organizations rely on? Knowing how and when to explore this unstructured data allows for future endeavors.

Innovation of Future Uses

There is no question that society is connected from being a part of retail memberships to businesses collecting inventory. Data surrounds a vast majority of our population. 95% of enterprises prioritize unstructured data management. There lies untapped information that companies have at their disposal that is unused. Recent projections reveal that unstructured data is over 80% of all enterprise data (2021, June 29). The prediction is that 175 billion zettabytes of unstructured data will increase by 2025 (Virtualitics, 2021). With this knowledge of projection, it is easy to see how data is a vital resource for future success. The future of analytics is machine learning (ML) and artificial intelligence (AI). It will enhance business intelligence and expand industrial innovation. Efficient usage of data formats and models will help businesses understand customer needs at a level of depth, focus on a targeted market, revise their products to offer more than competitors, and more (2021, June 29).

Conclusion

Obtaining the knowledge of unstructured data is vital no matter what business industry it is. Using unstructured data within a business setting is critical when viewing products. These projections are versatile in that they are not narrowed to one set department. Different departments include human resources, sales, marketing, and so on. Understanding how to analyze unstructured data and the tools available to dig into unstructured data opens opportunities that were not in the light before. Even with the few disadvantages of unstructured data, some companies can work on solutions so that those do not exist. Unstructured data can pave the way for future endeavors.

References

Actionable Tips to Analyze Unstructured Data. Michiganstateuniversityonline.com. (2019, May 28). Retrieved October 1, 2022, from https://www.michiganstateuniversityonline.com/resources/business-analytics/actionable-tips-to-analyze-unstructured-data/

Bavota, G. (2016, March). Mining unstructured data in software repositories: Current and future trends. In 2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER) (Vol. 5, pp. 1-12). IEEE.

Botha, M. (2019, September 19). *When to use qualitative vs. quantitative research*. Alchemer. Retrieved October 1, 2022, from https://www.alchemer.com/resources/blog/quantitative-qualitative-research/#:~:text=Qualitative%20research%20is%20by%20definition,to%20the%20problem%20at%20hand.

Chasupa, T. L., & Paireekreng, W. (2021, August). The Framework of Extracting Unstructured Usage for Big Data Platform. In 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP) (pp. 90-94). IEEE.

Davies, N. (2021, September 30). *Is unstructured data the future of data management?* DATAVERSITY. Retrieved October 1, 2022, from https://www.dataversity.net/is-unstructured-data-the-future-of-data-management/

Hansen, L. (2022, August 5). How businesses use unstructured data for bi. CIO Insight. Retrieved October 1, 2022, from https://www.cioinsight.com/it-strategy/bi-unstructured-data/

Ihritik. (2021, October 10). *What is unstructured data?* GeeksforGeeks. Retrieved October 1, 2022, from https://www.geeksforgeeks.org/what-is-unstructured-data/

Process Data Mining - Data Science Consulting Services: BTI. Business Transformation Institute. (2022, February 25). Retrieved October 1, 2022, from https://www.biztransform.net/process-data-mining/?gclid=Cj0KCQjwyt-ZBhCNARIsAKH1174htoZP2wbnlBtn5xtdpSFW88yL96MGHEtCjZPipu2kDktB_d5-xSUaAjJHEALw_wcB

Rose, C. (2021, May 20). *8 examples of unstructured data*. Treehouse Tech Group. Retrieved October 1, 2022, from https://treehousetechgroup.com/8-examples-of-unstructured-data/

Smith, T. (2022, August 17). *Qualitative analysis*. Investopedia. Retrieved October 1, 2022, from https://www.investopedia.com/terms/q/qualitativeanalysis.asp#:~:text=our%20editorial%20poli cies-,What%20Is%20Qualitative%20Analysis%3F,and%20development%2C%20and%20labor%20relations.

Structured vs. unstructured data: What's the difference? IBM. (2021, June 29). Retrieved October 1, 2022, from https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

Taylor, C. (2022, June 6). *Structured vs unstructured data 101: Top guide*. Datamation. Retrieved October 1, 2022, from https://www.datamation.com/big-data/structured-vs-unstructured-data/

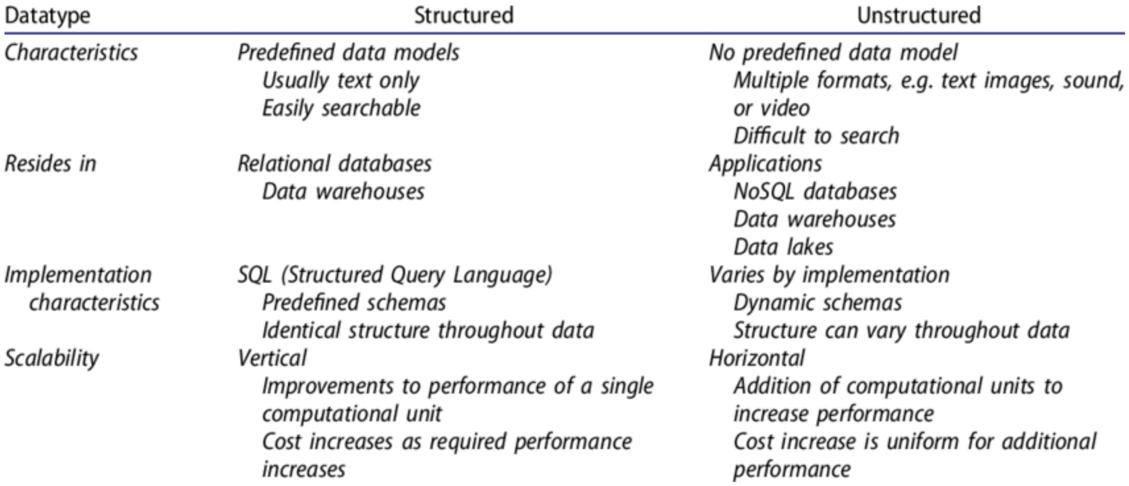
Virtualitics. (2021, December 16). *Is unstructured data the future of data management?* Is Unstructured Data the Future of Data Management? Retrieved October 1, 2022, from https://blog.virtualitics.com/is-unstructured-data-the-future-of-data-management

What is Big Data? Big Data explained. Tableau. (n.d.). Retrieved October 1, 2022, from https://www.tableau.com/learn/articles/what-is-big-data

What is unstructured data. Scion Analytics. (2021, July 8). Retrieved October 1, 2022, from https://scionanalytics.com/what-is-unstructured-data/

Appendix A

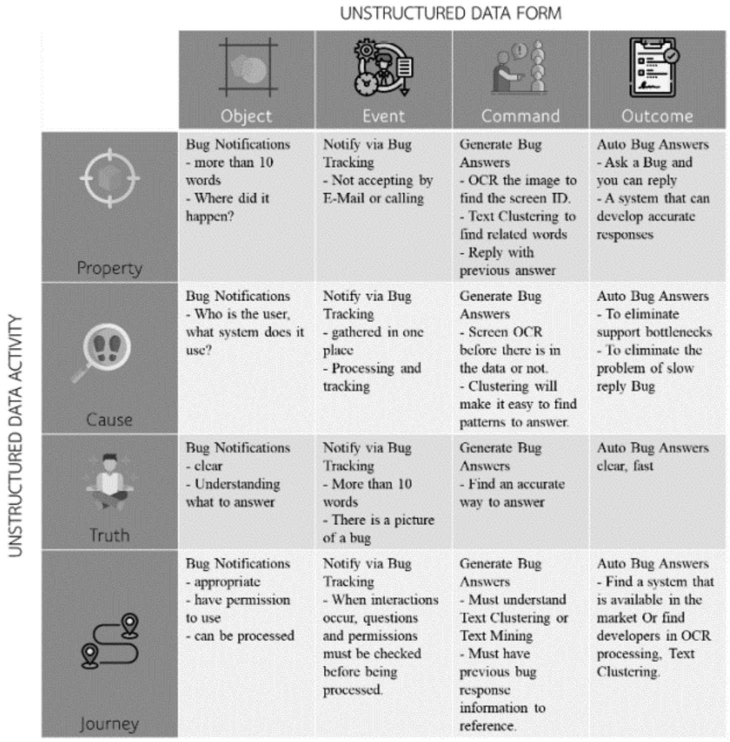
A table of how Structed and Unstructured Data compares



Datatype	Structured	Unstructured
Characteristics	*Predefined data models* *Usually text only* *Easily searchable*	*No predefined data model* *Multiple formats, e.g. text images, sound, or video* *Difficult to search*
Resides in	*Relational databases* *Data warehouses*	*Applications* *NoSQL databases* *Data warehouses* *Data lakes*
Implementation characteristics	*SQL (Structured Query Language)* *Predefined schemas* *Identical structure throughout data*	*Varies by implementation* *Dynamic schemas* *Structure can vary throughout data*
Scalability	*Vertical* *Improvements to performance of a single computational unit* *Cost increases as required performance increases*	*Horizontal* *Addition of computational units to increase performance* *Cost increase is uniform for additional performance*

Appendix B

A layout of unstructured data extracting model and trial run



UNSTRUCTURED DATA FORM

UNSTRUCTURED DATA ACTIVITY

	Object	Event	Command	Outcome
Property	Bug Notifications - more than 10 words - Where did it happen?	Notify via Bug Tracking - Not accepting by E-Mail or calling	Generate Bug Answers - OCR the image to find the screen ID. - Text Clustering to find related words - Reply with previous answer	Auto Bug Answers - Ask a Bug and you can reply - A system that can develop accurate responses
Cause	Bug Notifications - Who is the user, what system does it use?	Notify via Bug Tracking - gathered in one place - Processing and tracking	Generate Bug Answers - Screen OCR before there is in the data or not. - Clustering will make it easy to find patterns to answer.	Auto Bug Answers - To eliminate support bottlenecks - To eliminate the problem of slow reply Bug
Truth	Bug Notifications - clear - Understanding what to answer	Notify via Bug Tracking - More than 10 words - There is a picture of a bug	Generate Bug Answers - Find an accurate way to answer	Auto Bug Answers clear, fast
Journey	Bug Notifications - appropriate - have permission to use - can be processed	Notify via Bug Tracking - When interactions occur, questions and permissions must be checked before being processed.	Generate Bug Answers - Must understand Text Clustering or Text Mining - Must have previous bug response information to reference.	Auto Bug Answers - Find a system that is available in the market Or find developers in OCR processing, Text Clustering.

Figure 8. Unstructured Big Data Extracting Model and trial run.